

**A corpus of consonant-vowel-consonant (CVC) real words
and nonwords: Comparison of phonotactic probability,
neighborhood density, and consonant age-of-acquisition**

Journal:	<i>Behavior Research Methods</i>
Manuscript ID:	BR-Org-12-248.R1
Manuscript Type:	Original Manuscript
Date Submitted by the Author:	n/a
Complete List of Authors:	Storkel, Holly; University of Kansas, Speech-Language-Hearing
Note: The following files were submitted by the author for peer review, but marked to be sent in Off-Line.	
Excel file of CVCs	

Running Head: CVC Corpus

A corpus of consonant-vowel-consonant (CVC) real words and nonwords: Comparison of
phonotactic probability, neighborhood density, and consonant age-of-acquisition

Holly L. Storkel

University of Kansas

Revision Submitted, 11/29/2012, *Behavior Research Methods*

Corresponding author:

Holly Storkel, Ph.D.
Associate Professor
Department of Speech-Language-Hearing: Sciences and Disorders
University of Kansas
3001 Dole Human Development Center
1000 Sunnyside Avenue, Lawrence, KS 66045-7555
E-mail: hstorkel@ku.edu
Phone: 785-864-0497
Fax: 785-864-3974

Abstract

A corpus of 5,765 consonant-vowel-consonant (CVC) sequences was compiled, and phonotactic probability and neighborhood density based on both child and adult corpora were computed. This corpus of CVCs, provided as supplementary materials, was analyzed to address the following questions: (1) Do computations based on a child corpus differ from those based on an adult corpus? (2) Do phonotactic probability and/or neighborhood density of real words differ from that of nonwords? (3) Do phonotactic probability and/or neighborhood density differ across CVCs varying in consonant age-of-acquisition? Results showed significant differences in phonotactic probability and neighborhood density for child versus adult corpora, replicating prior findings. The impact of this difference on future studies will depend on the level of precision needed in specifying probability and density. In addition, significant and large differences in phonotactic probability and neighborhood density were detected between real words and nonwords, which may present methodological challenges for future research. Lastly, CVCs composed of earlier acquired sounds differed significantly in probability and density from CVCs composed of later acquired sounds, although this effect was relatively small and less likely to present significant methodological challenges to future studies.

(Word count: 187 of 250 words)

Key words: neighborhood density, phonotactic probability

A corpus of consonant-vowel-consonant (CVC) real words and nonwords: Comparison of phonotactic probability, neighborhood density, and consonant age-of-acquisition

Numerous studies show that *phonotactic probability*, the likelihood of occurrence of a sound sequence in a language, and *neighborhood density*, the number of words that are phonologically similar to a given sound sequence, influence spoken language recognition, production, and acquisition of both real words and nonwords across the lifespan (e.g., Munson, 2001; Munson, Swenson, & Manthei, 2005; Newman & German, 2005; Storkel, Armbruster, & Hogan, 2006; Storkel & Lee, 2011; Vitevitch & Luce, 1999). Given the clear influence of phonotactic probability and neighborhood density across multiple tasks, age groups, and types of stimuli (i.e., real words vs. nonwords) it is crucial to control or manipulate these variables in psycholinguistic research either during stimulus selection or data analysis. To support this, a corpus of CVCs was created (provided as supplemental materials); phonotactic probability and neighborhood density based on both child and adult corpora were measured; and potential relationships among CVCs were investigated to better inform stimulus selection. The specific issues addressed were whether (1) computations based on a child corpus differed from those based on an adult corpus; (2) phonotactic probability and/or neighborhood density of real words differed from that of nonwords; (3) phonotactic probability and/or neighborhood density differed across CVCs varying in consonant age-of-acquisition.

Comparison of Child and Adult Values

In terms of comparability of computations based on child versus adult corpora, a prior study by Storkel and Hoover (2010) addressed this issue for a set of 380 early acquired nouns that varied in word length and sound structure. Results showed that child values were

significantly correlated with adult values. However, the raw values did differ significantly with child phonotactic probability being higher than adult phonotactic probability and child neighborhood density being lower than adult neighborhood density. Transformation of values into z scores based on the means and standard deviations of the child or adult corpus reduced the difference between child and adult values. This finding indicates that significant differences in raw values were likely related to differences in the size and composition of the child versus the adult corpus, which were minimized by transformation of the values in a manner that is sensitive to the individual characteristics of the corpus. Similar findings were obtained for a non-random sample of 310 primarily CVC nonwords. The current report extends the issue of comparability of child and adult probability and density values to a large set of CVCs that includes both real words and nonwords. It is expected that the results of the prior study will be replicated, indicating the need to consider differences in corpora used to compute phonotactic probability and neighborhood density.

Lexicality and Consonant Age-of-Acquisition

Although Storkel and Hoover (2010) analyzed child and adult values for real words and nonwords, the two types of stimuli were never compared to one another. Thus, it is unclear whether the phonotactic probability or neighborhood density of real words differs from that of nonwords. Prior research suggests that the effect of phonotactic probability and neighborhood density may differ for real words versus nonwords (e.g., Munson, et al., 2005; Vitevitch, 2003; Vitevitch & Luce, 1998, 1999). In addition, phonotactic probability and neighborhood density are correlated with wordlikeness judgments (Bailey & Hahn, 2001; Frisch, Large, & Pisoni, 2000). That is, nonwords that are higher probability or higher density tend to be judged as sounding more like a real word than nonwords that are lower probability or lower density. It's

possible that this finding could be further extended to show that real words are higher probability and/or higher density than nonwords. An understanding of how phonotactic probability and neighborhood density vary by lexicality may inform stimulus selection for future research.

In a similar vein, past research indicates that phonotactic probability and neighborhood density can influence accuracy of sound production with production generally being more accurate for high probability and/or high density sound sequences (e.g., Edwards, Beckman, & Munson, 2004; Gierut & Storkel, 2002; Vitevitch, 1997; Zamuner, Gerken, & Hammond, 2004). Moreover, it has been argued that phonological acquisition in children is tightly coupled with acquisition and knowledge of words (Edwards, Munson, & Beckman, 2011; Stoel-Gammon, 2011; Velleman & Vihman, 2002). One question that arises is whether CVCs composed of earlier acquired sounds might have higher phonotactic probability and/or neighborhood density than CVCs composed of later acquired sounds, a finding that would be informative for designing developmental studies of phonotactic probability or neighborhood density.

Purpose

The purpose of the current report is to provide a comprehensive corpus of legal CVCs in American English (see supplemental materials) that can be used in psycholinguistic research. To that end, phonotactic probability and neighborhood density are computed based on child and adult corpora, and CVCs are coded as real words or nonwords and by consonant age-of-acquisition. Three questions are addressed: (1) Do phonotactic probability and/or neighborhood density values differ depending on the corpus (i.e., child vs. adult) used for the computations? (2) Are real word CVCs higher in phonotactic probability and/or neighborhood density than nonword CVCs? (3) Are CVCs composed of earlier acquired sounds higher in phonotactic probability and/or neighborhood density than CVCs composed of later acquired sounds?

Method

Child and Adult Corpora

Variables of interest were determined using an online calculator available at http://www.bncdnet.ku.edu/cml/info_ccc.vi. The child corpus for this calculator is described more fully in Storkel and Hoover (2010). In short, this corpus consists of 4,832 different words spoken by American Kindergarten or first grade children (Kolson, 1960; Moe, Hopkins, & Rush, 1982). The adult corpus is described more fully in (Nusbaum, Pisoni, & Davis, 1984). Briefly, this corpus consists of 19,290 words taken from a dictionary of American English (*The new Merriam-Webster pocket dictionary*, 1964). For each word, both corpora contain a phonetic transcription of the target pronunciation in American English in a computer readable format, an orthographic spelling of the word, and the log frequency of the word based on a sample of approximately 1 million words. Both corpora generally consist of uninflected root words (e.g., “run” rather than “running”) because this is the typical format for dictionaries, and the child corpus was created to match this format (see Storkel & Hoover, 2010 for details).

Lexicality

Consonant-vowel-consonant sequences were generated by pairing all possible combinations of initial consonants, vowels, and final consonants. These CVCs were then submitted to the online calculator (i.e., http://www.bncdnet.ku.edu/cml/info_ccc.vi) that computes phonotactic probability and neighborhood density based on the child (Storkel & Hoover, 2010) or adult (Nusbaum, et al., 1984) corpus. Importantly, the calculator also identifies whether the input item occurs in either corpus. In this way, real word CVCs were differentiated from nonword CVCs. A real word was defined as any CVC that occurred in the child corpus only ($n = 84$), the adult corpus only ($n = 592$), or both corpora ($n = 720$). Thus, 1,396 CVCs were

identified as real words. For the remaining CVCs, which form the pool of potential nonwords, those with phonotactic probability and neighborhood density equal to zero were removed from consideration because these were considered to be unattested sequences in this sample of American English. This yielded 4,369 CVCs identified as probable nonwords meeting the characteristics of American English. The Excel file provided in the supplemental materials has three worksheets showing (1) real word CVCs; (2) nonword CVCs; (3) real and nonword CVCs combined. The data in the last worksheet (i.e., real and nonword CVCs combined) were analyzed for this report.

Phonotactic Probability

Two raw measures of phonotactic probability were computed based on each corpus (child vs. adult), using the online calculator: (1) positional segment sum; (2) biphone sum. *Positional segment sum* is computed by first calculating the positional segment frequency for each sound in the CVC and then adding those individual frequencies together. *Positional segment frequency* is computed by summing the log frequency of all the words in a corpus that contain the given sound in the given word position and then dividing by the sum of the log frequency of all the words in the corpus that contain any sound in the same word position. *Biphone sum* is computed in a similar manner but the unit of calculation is the pair of adjacent sounds (i.e., CV or VC), rather than a single sound. Thus, *biphone frequency* for a given pair of sounds is the sum of the log frequency of all the words in the corpus that contain the given sound *pair* in the given word position divided by the sum of the log frequency of all the words in the corpus that contain any sound in the given word position. Storkel (2004b) provides a detailed example of these calculations.

1
2
3 In addition to these raw values, transformed values were computed. The transformations
4
5 were computed for real words alone, nonwords alone, and real words and nonwords combined.
6
7 For each of these three sets, the mean and standard deviation for each measure of phonotactic
8
9 probability (i.e., positional segment and biphone sums) was computed for each corpus (i.e., child
10
11 and adult) and then used to compute a z score and percentile for each CVC. The formula for the z
12
13 score is: (obtained value – mean)/standard deviation. Percentiles were computed using an SPSS
14
15 function (i.e., cdfnorm) that computes the percentile based on a normal curve with the given
16
17 mean and standard deviation. The z scores for the real words and nonwords combined are the
18
19 data that were analyzed for this report so that real words could be compared to nonwords. Table
20
21 1 provides the means and standard deviations used to create these z scores.
22
23
24
25
26

27 Note that the raw value for a given CVC is the same across all worksheets in the
28
29 supplemental Excel file but the transformed value changes across worksheets because the mean
30
31 and standard deviation used for the transformation is specific to a given worksheet (i.e., set of
32
33 CVCs: real words only, nonwords only, both). For future studies using the supplemental
34
35 materials to select stimuli, a particular worksheet should be chosen based on the correspondence
36
37 with the type of stimuli needed for the study. For example, if only real word CVCs are being
38
39 used in the study, then the real word worksheet should be used for stimuli selection; whereas, if
40
41 real word and nonword CVCs are being used, then the all CVC worksheet should be used for
42
43 stimuli selection. The transformed values indicate how extreme a particular CVC is relative to
44
45 the other CVCs in the same set/worksheet. That is, for the first case of real words only, a z score
46
47 of +1.0 for positional segment sum indicates that the selected CVC has a positional segment sum
48
49 that is 1.0 standard deviations above the mean positional segment sum of *real word* CVCs;
50
51
52
53
54
55
56
57
58
59
60 whereas, in the second case of real words and nonwords, a z score for positional segment sum of

+1.0 indicates that the selected CVC’s positional segment sum is 1.0 standard deviations above the mean positional segment sum for *all* CVCs. Lastly, *z* scores place the positional segment sum and biphone sum on the same scale, making it appropriate to average the two *z* scores to create one measure of phonotactic probability when a single measure is desirable.

Neighborhood Density

Neighborhood density was computed for each corpus (child or adult) by counting the number of words appearing in the corpus that differed from the given CVC by a one sound substitution, deletion, or addition in any word position. To illustrate, the neighbors of the real word CVC “rat” include “bat” (initial sound substitution), “rot” (middle sound substitution), “rag” (final sound substitution), “at” (initial sound deletion), “brat” (initial sound addition), and “raft” (final sound addition). Note that determination of neighbors is based on sounds, rather than spelling. As with phonotactic probability, transformed values, specifically *z* scores and percentiles, were computed for neighborhood density following the methods already described. See Table 1 for the means and standard deviations used for these transformations.

Consonant Age-of-Acquisition

Categories of consonant age-of-acquisition were taken from Shriberg (1993) who divided the 24 American English consonants into three groups of 8 consonants based on accuracy by a group of children with speech sound disorders. The groupings identified by Shriberg were consistent with data from larger cross-sectional studies of typically developing children (e.g., Smit, Hand, Freilinger, Bernthal, & Bird, 1990). The three groupings are: (1) early-8, consisting of the sounds m, n, w, y, h, p, b, d; (2) middle-8, consisting of ng (e.g., king), t, k, g, f, v, ch, j; (3) late-8, consisting of voiceless th (e.g., thanks), voiced th (e.g., that), s, z, sh, zh (e.g., azure), l, r. Each consonant (initial and final) in a CVC was coded as early-, middle-, or late-8, and then

each CVC was given a whole-CVC code based on the coding of the two consonants. There were five whole-CVC codes. Specifically, code 1 (early) was assigned to CVCs where both consonants were early-8 sounds ($n = 567$). Code 2 (early/mid) was assigned to CVCs with one early-8 and one middle-8 sound ($n = 1387$). Code 3 (mid) was assigned to CVCs with one early-8 and one late-8 sound and to CVCs with two middle-8 sounds ($n = 1938$). Code 4 (mid/late) was assigned to CVCs with one middle-8 and one late-8 sound ($n = 1314$). Lastly, code 5 (late) was assigned to CVCs where both consonants were late-8 sounds ($n = 559$).

Results

All analyses were performed on the combined real word and nonword set of CVCs (i.e., All CVCs worksheet in the supplemental materials).

Comparison of Child and Adult Values

Comparison of child and adult values mirrored the findings of the prior study examining a different set of real words and nonwords (Storkel & Hoover, 2010). Figure 1 shows the child and adult raw values for positional segment sum (top panel), biphone sum (middle panel), and neighborhood density (bottom panel). As shown in Figure 1, raw positional segment sum, biphone sum, and neighborhood density based on the child corpus were significantly correlated with raw values based on the adult corpus, $r(5765) = .95, p < .001, r^2 = .91$ for positional segment sum; $r(5765) = .88, p < .001, r^2 = .78$ for biphone sum, $r(5765) = .89, p < .001, r^2 = .79$ for neighborhood density. However, t test analysis showed that positional segment sum based on the child corpus was significantly higher than that based on the adult corpus, $t(5764) = -61.81, p < .001$. Likewise, biphone sum based on the child corpus was significantly higher than that based on the adult corpus, $t(5764) = -23.93, p < .001$. In contrast, neighborhood density based on the child corpus was significantly lower than that based on the adult corpus, $t(5764) =$

104.77, $p < .001$. Because of this significant difference, z scores were used in subsequent analyses to re-scale the measures on a common metric and minimize differences across corpora (Storkel & Hoover, 2010).

Lexicality and Consonant Age-of-Acquisition

Three separate multivariate analyses of variance (MANOVA) were performed: one for each dependent variable (i.e., positional segment sum, biphone sum, and neighborhood density z scores). In each MANOVA, lexicality (real word vs. nonword) and whole-CVC consonant age-of-acquisition (1-early, 2-early/mid, 3-mid, 4-mid/late, 5-late) were the independent variables, and z score based on the child and adult corpus were the dependent variables. MANOVA was used because the dependent variables based on the child and adult corpora were correlated, making a univariate approach inappropriate due to inflation of Type I and Type II error rates (e.g., Haase & Ellis, 1987). However, it is important to note that power in MANOVA is affected by the correlation between the dependent variables, such that more highly correlated dependent variables, as in the current report, will tend to reduce power (Cole, Maxwell, Arvey, & Salas, 1994). Thus, the analyses reported here represent a potentially conservative analysis approach, although the relatively large sample size ($n = 5,765$) offsets this possible limitation.

For positional segment sum, the effect of lexicality was significant, $F(2, 5754) = 281.91$, $p < .001$, Wilks' $\lambda = .91$, $\eta_p^2 = 0.10$. As shown in the top panel of Figure 2, real words had higher positional segment sums than nonwords, and this was true for the child, $F(1, 5755) = 563.00$, $p < .001$, $\eta_p^2 = 0.09$, and the adult corpus, $F(1, 5755) = 489.33$, $p < .001$, $\eta_p^2 = 0.08$. Likewise, the effect of consonant age-of-acquisition was significant, $F(8, 11508) = 11.85$, $p < .001$, Wilks' $\lambda = .98$, $\eta_p^2 = 0.01$. The effect was significant for both the child, $F(4, 5755) = 18.04$, $p < .001$, $\eta_p^2 = 0.01$, and the adult corpus, $F(4, 5755) = 11.98$, $p < .001$, $\eta_p^2 = 0.01$. This significant effect was

further examined via Tukey HSD. As shown in the top panel of Figure 2, CVCs with two early consonants (i.e., 1-early) had significantly higher positional segment sums than all other combinations of consonant age-of-acquisition, all $ps < .001$ for child and adult corpora. In contrast, CVCs with one middle and one late consonant (i.e., 4-mid/late) had significantly lower positional segment sums than all other combinations of consonant age-of-acquisition, all $ps < .01$ for child and adult corpora. As can be seen in the top panel of Figure 2, lexicality did not significantly interact with consonant age-of-acquisition, $F(8, 11508) = 1.13, p = .34$, Wilks' $\lambda = .998, \eta_p^2 = 0.001$.

For biphone sum, the effect of lexicality was significant, $F(2, 5754) = 411.09, p < .001$, Wilks' $\lambda = .88, \eta_p^2 = 0.13$. As shown in the middle panel of Figure 2, real words had higher biphone sums than nonwords, and this was true for the child, $F(1, 5755) = 808.80, p < .001, \eta_p^2 = 0.12$, and the adult corpus, $F(1, 5755) = 517.99, p < .001, \eta_p^2 = 0.08$. Likewise, the effect of consonant age-of-acquisition was significant, $F(8, 11508) = 9.96, p < .001$, Wilks' $\lambda = .99, \eta_p^2 = 0.01$. The effect was significant for both the child, $F(4, 5755) = 13.40, p < .001, \eta_p^2 = 0.01$, and the adult corpus, $F(4, 5755) = 6.23, p < .001, \eta_p^2 = 0.004$. As shown in the middle panel of Figure 2, CVCs with two early acquired consonants (i.e., 1-early) had significantly higher biphone sums than all other combinations of consonant age-of-acquisition, all $ps < .01$ for child and adult corpora. In contrast, CVCs with one middle and one late acquired consonant (i.e., 4-mid/late) had lower biphone sums than most other combinations of consonant age-of-acquisition, all $ps < .05$ for the child (except 5-late/late) and adult corpus (except 2-early/mid). As shown in the middle panel of Figure 2, lexicality did not significantly interact with consonant age-of-acquisition, $F(8, 11508) = 0.29, p = .97$, Wilks' $\lambda = 1.00, \eta_p^2 < 0.001$.

For neighborhood density, the effect of lexicality was significant, $F(2, 5754) = 728.26, p < .001$, Wilks' $\lambda = .80, \eta_p^2 = 0.20$. As shown in the bottom panel of Figure 2, real words had higher densities than nonwords, and this was true for the child, $F(1, 5755) = 1182.77, p < .001, \eta_p^2 = 0.17$, and the adult corpus, $F(1, 5755) = 1446.61, p < .001, \eta_p^2 = 0.20$. Likewise, the effect of consonant age-of-acquisition was significant, $F(8, 11508) = 38.89, p < .001$, Wilks' $\lambda = .95, \eta_p^2 = 0.03$. The effect was significant for both the child, $F(4, 5755) = 42.38, p < .001, \eta_p^2 = 0.03$, and the adult corpus, $F(4, 5755) = 10.72, p < .001, \eta_p^2 = 0.01$. As shown in the bottom panel of Figure 2, density tended to decrease as consonant age-of-acquisition increased. All pairwise comparisons were significant for the child corpus, all $ps < .05$, and most pairwise comparisons were significant for the adult corpus, all $ps < .01$ except 2-early/mid vs. 3-mid and 4-mid-late vs. 5-late-late. These significant main effects were qualified by a significant interaction between lexicality and consonant age-of-acquisition, $F(8, 11508) = 7.91, p < .001$, Wilks' $\lambda = .99, \eta_p^2 = 0.01$. Note that the interaction was significant only for the adult corpus, $F(4, 5755) = 4.69, p = .001, \eta_p^2 = 0.003$, and not the child corpus, $F(4, 5755) = 0.32, p = .87, \eta_p^2 < 0.001$. As shown in the bottom panel of Figure 2, this interaction appeared to be attributable to a stronger effect of consonant age-of-acquisition on density for nonwords rather than real words, especially for density based on the adult corpus. That is, the effect of consonant age-of-acquisition on density was significant for nonwords, $F(4, 4364) = 50.21, p < .001, \eta_p^2 = 0.04$, and real words, $F(4, 1391) = 9.50, p < .001, \eta_p^2 = 0.03$, for the child corpus, but the effect was only significant for nonwords, $F(4, 4364) = 26.00, p < .001, \eta_p^2 = 0.02$, and not real words, $F(4, 1391) = 1.35, p = .25, \eta_p^2 < 0.01$, for the adult corpus.

Discussion

To summarize, the three main findings are that (1) phonotactic probability based on a child corpus was higher than that based on an adult corpus whereas neighborhood density based on the child corpus was lower than that based on the adult corpus; (2) real word CVCs were higher probability and higher density than nonword CVCs; (3) CVCs composed of earlier acquired sounds were higher probability and higher density than CVCs composed of later acquired sounds. These three findings have both theoretical and methodological implications.

Comparison of Child and Adult Values

The first finding replicates a prior study using the same child and adult corpora and calculator (Storkel & Hoover, 2010) but with a larger and more varied set of CVCs. However, the prior explanation of these significant differences across corpora is likely relevant to the current findings. Specifically, the prior analysis of these two corpora showed that the words in the child corpus were higher in frequency than the words in the adult corpus (Storkel & Hoover, 2010), possibly as a by-product of frequency effects on learning. That is, a child's lexicon is likely to consist predominately of high frequency words, which are easier to learn (e.g., Storkel, 2004a). As the lexicon grows, low frequency words are added, such that the adult lexicon consists of a mix of low and high frequency words. Because word frequency is used in calculating phonotactic probability, this change in the frequency of the words in the lexicon could account for the observed lowering of phonotactic probability from the child to the adult corpus. In complement, as words are added to the lexicon, the overall size of the lexicon changes, including the size of individual neighborhoods (e.g., Charles-Luce & Luce, 1990, 1995). The observed increase in neighborhood density from the child to the adult corpus is consistent with these prior results.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

In terms of methodological implications, it's important to note that child phonotactic probability and neighborhood density were highly correlated with adult phonotactic probability or neighborhood density. Thus, when only broad distinctions (e.g., low vs. high probability or density) are being studied across ages, it will likely be possible to identify stimuli that are low or high for both child and adult measures of probability or density. However, studying finer distinctions in phonotactic probability or neighborhood density may be more challenging because of the differences across corpora that likely also reflect differences across age (e.g., changes in the size of the corpus likely mirror changes in the size of the lexicon). To illustrate, a density of 5 neighbors may not have the same "meaning" across the child and the adult lexicon. Specifically, 5 neighbors is relatively close to the mean density for children (i.e., z score = -0.62) but relatively farther from the mean density for adults (i.e., z score = -1.11). Moreover, it is unclear whether it is the raw density or the relative density that critically influences language processing. That is, does the presence of 5 neighbors have the same effect on language processing regardless of where this falls in the density distribution (i.e., raw values matter) or does the words relative position within the system (i.e., the degree of sparseness) influence processing (i.e., relative measures matter)? Note that a similar scenario could be constructed for phonotactic probability. In selecting stimuli for developmental studies investigating finer distinctions of phonotactic probability and neighborhood density, the theoretical framework would need to be considered to determine whether raw or relative values are predicted to influence language processing. If strong predictions are not possible, then both types of values may need to be investigated to determine which aspect of probability or density influences language processing.

Lexicality

The finding that real words are higher probability and higher density than nonwords is consistent with prior studies of wordlikeness ratings, where higher probability and higher density nonwords are judged as more wordlike than lower probability or lower density nonwords (Bailey & Hahn, 2001; Frisch, et al., 2000). That is, there appears to be a relationship between lexicality or potential lexicality (i.e., wordlikeness) and phonotactic probability and neighborhood density, such that higher probability and higher density CVCs are preferred. Thus, across the distribution of legal CVCs, the most probable and dense CVCs tend to be actual words in the language whereas the least probable and dense CVCs tend to be excluded from the language. This fits well with other studies, suggesting that language growth (i.e., adding new words to a language) is governed by *preferential attachment* (e.g., Perc, 2012), a process by which new items that are highly similar to existing items are more likely to be added to a system than new items that are less similar.

Turning to methodological considerations, it is particularly notable that the difference between nonword and real word CVCs was quite strong with relatively large effect sizes (i.e., η_p^2 range of 0.10 – 0.20). Moreover, z score differences were approximately 1.00 for most comparisons (see Figure 2), meaning that values for real words and nonwords differed by approximately one standard deviation (see Figure 2). The practical implication of this is that matching real words and nonwords on phonotactic probability and/or neighborhood density requires careful attention during stimuli selection and may not be possible, depending on other study-specific criteria. If matching is not possible, then there may be challenges in interpreting effects of lexicality, phonotactic probability, and neighborhood density. For example, if lexicality is manipulated without specific attention to phonotactic probability and neighborhood

density, it is likely that the items differ in probability and density, which could serve to amplify or dampen the effect of lexicality. Thus, results could not be solely attributable to lexicality. In this scenario, more complex statistical analyses (e.g., crossed-random effects multi-level modeling) might be a useful post-hoc solution, providing a means to account for differences in phonotactic probability and/or neighborhood density within the statistical analysis. Whether this is a viable solution would depend on precisely how large the probability and/or density difference is between the real words and nonwords. In addition, when comparing effects of phonotactic probability and neighborhood density across real words and nonwords, there could be two potential interpretations. The first is that phonotactic probability and/or neighborhood density affects processing of real words differently than processing of nonwords. The second is that phonotactic probability and/or neighborhood density affects processing differently at different points in the probability and/or density distribution (cf., Storkel, Bontempo, Aschenbrenner, Maekawa, & Lee, In Review, for a study showing differences in word learning across the full distribution of probability and density). Again, statistical analyses may be helpful in ruling out or supporting one of these alternatives over the other.

Consonant Age-of-Acquisition

Turning to consonant age-of-acquisition, the finding that CVCs composed of earlier acquired sounds have higher probability and higher density than CVCs composed of later acquired sounds is consistent with claims that phonological and lexical development are tightly coupled (Edwards, et al., 2011; Stoel-Gammon, 2011; Velleman & Vihman, 2002). In this case, both sound and word characteristics converge on favorable characteristics that should facilitate correct production.

In terms of methodological implications, the phonotactic probability and neighborhood density difference between CVCs composed of earlier acquired sounds and CVCs composed of later acquired sounds was somewhat weak with relatively small effect sizes (i.e., η_p^2 range of 0.01 – 0.03) and z score differences of approximately 0.50 or less (see Figure 2). Unlike the lexicality effect on phonotactic probability and neighborhood density, the consonant age-of-acquisition effect should be relatively easier to contend with when designing a study because there is a fair degree of overlap in the probability and density distributions for each consonant age-of-acquisition category (see Figure 2). Thus, it is likely that CVCs composed of earlier acquired sounds could be closely matched in phonotactic probability and/or neighborhood density to CVCs composed of later acquired sounds, isolating consonant age-of-acquisition effects in an empirical study. Likewise, it should be possible to define ranges for low versus high phonotactic probability and/or neighborhood density that are the same for CVCs composed of earlier versus later acquired sounds, making it possible to cleanly cross phonotactic probability and/or neighborhood density with consonant age-of-acquisition.

Conclusions

Child and adult corpora yield differing values for phonotactic probability and neighborhood density, but the implication of this for future research will depend on the level of precision needed in manipulating probability or density. In contrast, large differences in phonotactic probability and neighborhood density exist between real words and nonwords, which may present methodological challenges in designing studies manipulating lexicality and phonotactic probability or neighborhood density. Although CVCs composed of earlier acquired sounds did differ in probability and density from CVCs composed of later acquired sounds, this effect was relatively small and less likely to present significant methodological challenges to

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

studies manipulating consonant age-of-acquisition and phonotactic probability or neighborhood density.

Author Note

Holly L. Storkel, Department of Speech-Language-Hearing: Sciences and Disorders,
University of Kansas.

The project described was supported by grant DC 08095 from NIH. The contents are
solely the responsibility of the authors and do not necessarily represent the official views of the
NIH.

Contact author: Holly Storkel, Ph.D., Associate Professor, Department of Speech-
Language-Hearing: Sciences and Disorders, University of Kansas, 3001 Dole Human
Development Center, 1000 Sunnyside Avenue, Lawrence, KS 66045-7555. E-mail:
hstorkel@ku.edu.

References

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44(4), 568-591.

Charles-Luce, J., & Luce, P. A. (1990). Similarity neighbourhoods of words in young children's lexicons. *Journal of Child Language*, 17(1), 205-215.

Charles-Luce, J., & Luce, P. A. (1995). An examination of similarity neighbourhoods in young children's receptive vocabularies. *Journal of Child Language*, 22(3), 727-735.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Quantitative Methods in Psychology*, 115, 465-474.

Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research*, 47(2), 421-436.

Edwards, J., Munson, B., & Beckman, M. E. (2011). Lexicon-phonology relationships and dynamics of early language development—A commentary on Stoel-Gammon's 'Relationships between lexical and phonological development in young children.'. *Journal of Child Language*, 38(1), 35-40.

Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language*, 42(4), 481-496.

Gierut, J. A., & Storkel, H. L. (2002). Markedness and the grammar in lexical diffusion of fricatives. *Clinical Linguistics & Phonetics*, 16(2), 115-134.

- 1
2
3 Haase, R. F., & Ellis, M. V. (1987). Multivariate analysis of variance. *Journal of Counseling*
4
5 *Psychology*, 34, 404-413.
6
7
8 Kolson, C. J. (1960). *The vocabulary of kindergarten children*. Unpublished Doctoral
9
10 Dissertation, University of Pittsburgh, Pittsburgh.
11
12
13 Moe, A. J., Hopkins, K. J., & Rush, R. T. (1982). *The vocabulary of first grade children*.
14
15 Springfield, IL: Thomas.
16
17
18 Munson, B. (2001). Phonological pattern frequency and speech production in adults and
19
20 children. *Journal of Speech, Language, and Hearing Research*, 44, 778-792.
21
22
23 Munson, B., Swenson, C. L., & Manthei, S. C. (2005). Lexical and Phonological Organization in
24
25 Children: Evidence from Repetition Tasks. *Journal of Speech, Language, and Hearing*
26
27 *Research*, 48(1), 108-124.
28
29
30 *The new Merriam-Webster pocket dictionary*. (1964). New York: Merriam-Webster.
31
32 Newman, R. S., & German, D. J. (2005). Life span effects of lexical factors on oral naming.
33
34 *Language and Speech*, 48(2), 123-156.
35
36
37 Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). Sizing up the Hoosier mental lexicon
38
39 *Research on Spoken Language Processing Report No. 10* (pp. 357-376). Bloomington,
40
41 IN: Speech Research Laboratory, Indiana University.
42
43
44 Perc, M. (2012). Evolution of the most common English words and phrases over the centuries.
45
46 *Journal of the Royal Society Interface*, 9, 3323-3328.
47
48
49 Shriberg, L. (1993). Four new speech and prosody-voice measures for genetics research and
50
51 other studies in developmental phonological disorders. *Journal of Speech and Hearing*
52
53 *Research*, 36, 105-140.
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E., & Bird, A. (1990). The Iowa Articulation Norms Project and its Nebraska replication. *Journal of Speech and Hearing Disorders*, 55(4), 779-798.

Stoel-Gammon, C. (2011). Relationships between lexical and phonological development in young children. *Journal of Child Language*, 38(1), 1-34.

Storkel, H. L. (2004a). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, 25(2), 201-221.

Storkel, H. L. (2004b). Methods for minimizing the confounding effects of word length in the analysis of phonotactic probability and neighborhood density. *Journal of Speech, Language, and Hearing Research*, 47(6), 1454-1468.

Storkel, H. L., Armbruster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*, 49(6), 1175-1192.

Storkel, H. L., Bontempo, D. E., Aschenbrenner, A. J., Maekawa, J., & Lee, S. Y. (In Review). The effect of incremental changes in phonotactic probability and neighborhood density on word learning by preschool children. *Manuscript submitted for publication*.

Storkel, H. L., & Hoover, J. R. (2010). An on-line calculator to compute phonotactic probability and neighborhood density based on child corpora of spoken American English. *Behavior Research Methods*, 42, 497-506.

Storkel, H. L., & Lee, S. Y. (2011). The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, 26(2), 191-211.

Velleman, S., & Vihman, M. (2002). Whole-Word Phonology and Templates: Trap, Bootstrap, or Some of Each? *Language, Speech, and Hearing Services in Schools*, 33, 9-23.

Vitevitch, M. S. (1997). The neighborhood characteristics of malapropisms. *Language and Speech*, 40, 211-228.

Vitevitch, M. S. (2003). The influence of sublexical and lexical representations on the processing of spoken words in English. *Clinical Linguistics & Phonetics*, 17(6), 487-499.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science*, 9, 325-329.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory of Language*, 40, 374-408.

Zamuner, T. S., Gerken, L., & Hammond, M. (2004). Phonotactic probabilities in young children's speech production. *Journal of Child Language*, 31, 515-536.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Table 1. Means and standard deviations used for z score transformations

		Child Corpus	Adult Corpus
Positional Segment Sum	<i>M</i>	0.1294	0.1176
	<i>SD</i>	0.0470	0.0464
Biphone Sum	<i>M</i>	0.0046	0.0040
	<i>SD</i>	0.0041	0.0041
Neighborhood Density	<i>M</i>	8.3	13.4
	<i>SD</i>	5.3	7.6

Note. Ranges are provided in the supplemental materials

Figure Captions

Figure 1. Scatter plots of child versus adult positional segment sum (top), biphone sum (middle), and neighborhood density (bottom). Solid line indicates the linear regression fit line. Dashed line is a reference line indicating a perfect correlation.

Figure 2. Normalized (i.e., z-score) positional segment sum (top), biphone sum (middle), and neighborhood density (bottom) by consonant acquisition class for real words based on the adult (open bar) or child corpus (vertical line bar) and nonwords based on the adult (filled bar) or child corpus (dotted bar). Error bars indicate standard errors.

Figure 1

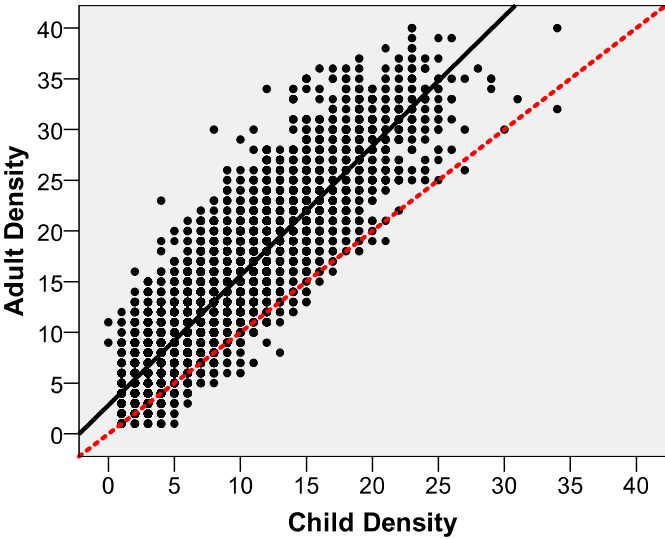
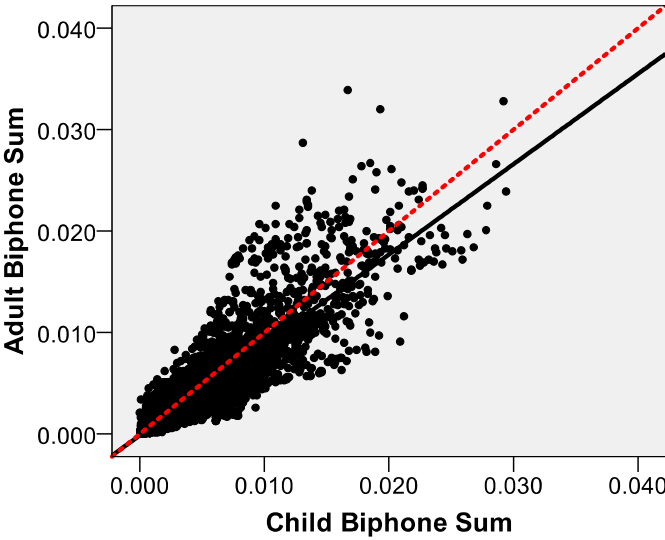
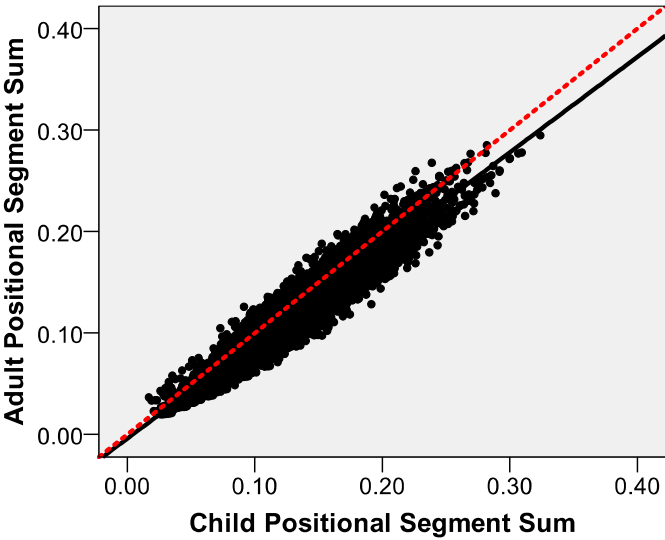


Figure 2

